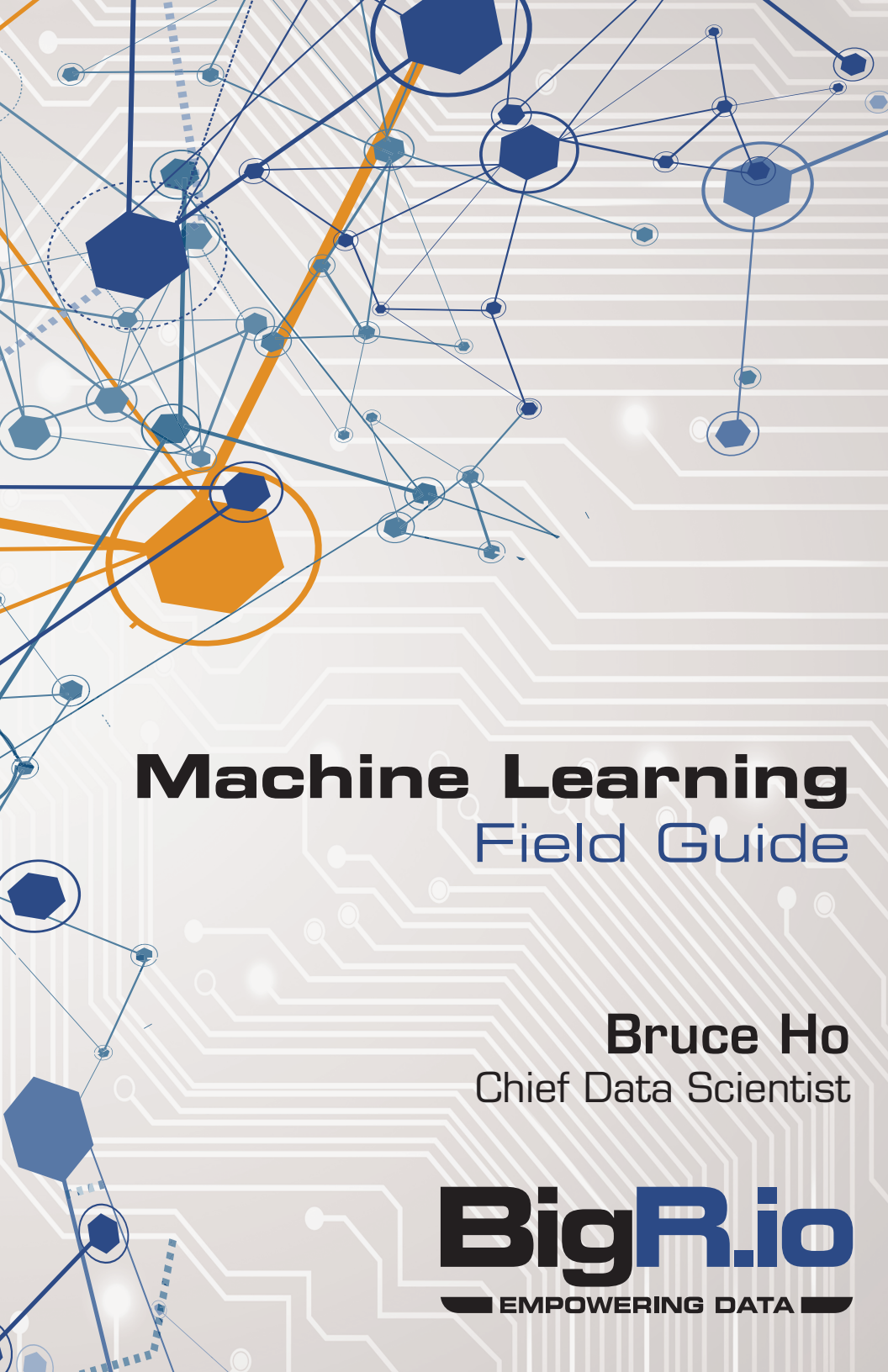


Machine Learning
Field Guide

Bruce Ho
Chief Data Scientist

BigR.io
EMPOWERING DATA

Machine Learning Field Guide

by Bruce Ho, Ph.D.

BigR.io Chief Data Scientist

BigR.io
EMPOWERING DATA

Introduction

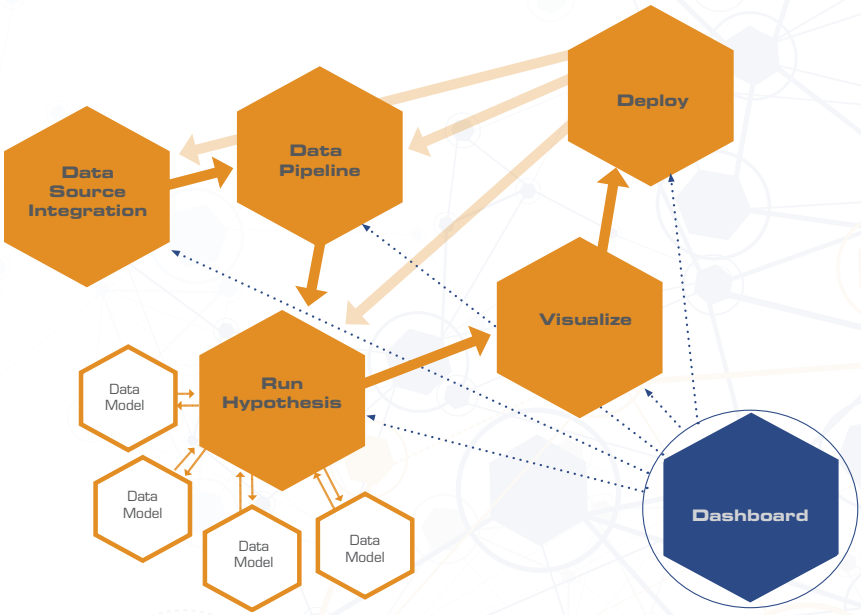
2017 has been named the Year of Machine Learning by tech visionaries everywhere. Businesses from eCommerce, to Healthcare, to Manufacturing, to Life Sciences, to Finance are feeling the inevitability of this latest digital transformation.

Boosted by cloud computing, Big Data, and advances in deep learning, a data-driven management philosophy is quickly becoming the norm. In the automobile industry, for example, it's projected that by 2020, 75% of cars will be connected generating 30,000 sensor signals per car (Gartner). Intelligence gathered from such data can reduce non-alcohol related accidents by 80% (US Department of Transportation).

How does a traditional enterprise keep up with the trend, stay relevant, and even lead the pack in embracing this new world where daily decisions center around predictive modeling and advanced data visualization?

The following process illustrates the concepts necessary to take on this initiative.

The Machine Learning Workflow



BigR.io
EMPOWERING DATA

Almost all enterprises are sitting on a goldmine of data and collecting more at a staggering rate and with ever greater complexity. Machine learning is about mining this treasure trove, extracting actionable business insights, predicting future events, and prescribing next best actions (NBAs), to achieve business goals. It starts with applying Big Data skills to integrate all available data into a data lake, and then cleansing and sampling the right data to feed into the predictive model for training.

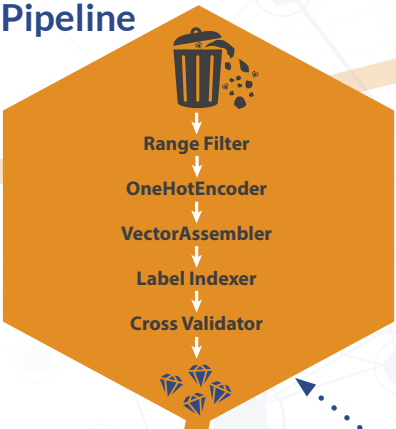
A manageable training procedure is designed by selecting from over a hundred well known statistical algorithms, which in current day practice include deep neural network techniques. The training process is still considered high art, relying on a constrained pool of talent well versed in advanced mathmatical algorithms. The distillation of the elaborate training process is a condensed set of model parameters which capture the optimal formula capable of predicting business outcomes with remarkable levels of accuracy.

The model is immersed into the operational environment, where front line staff conduct their day-to-day business with this artificial intelligence crystal ball.

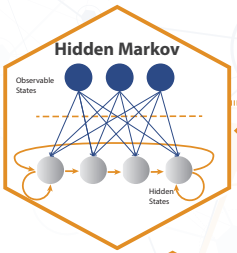
The Machine Learning Workflow

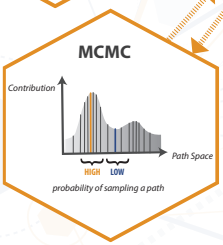
1. Data Source Integration

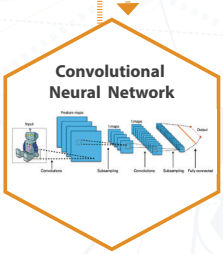
2. Data Pipeline



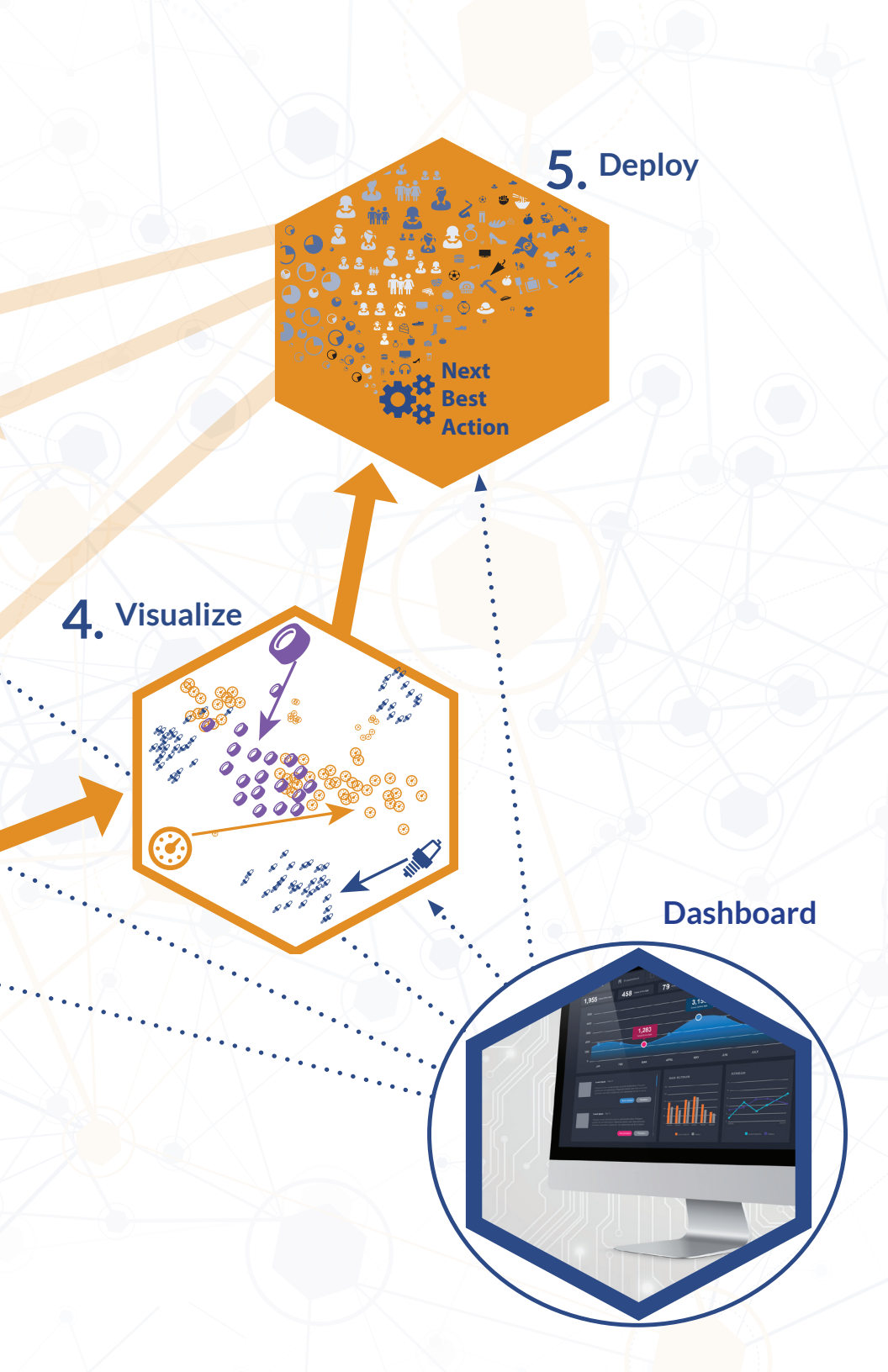
3. Run Hypothesis

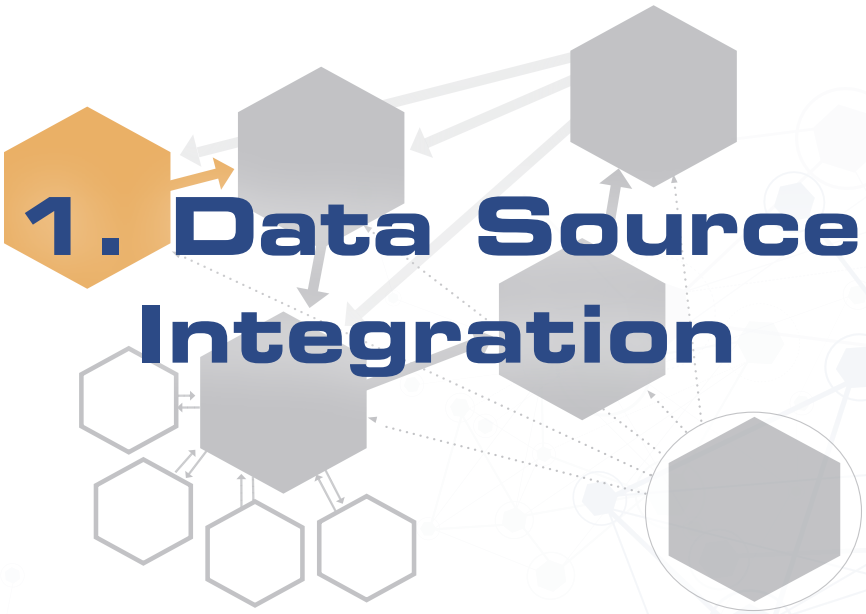






BigR.io
EMPOWERING DATA


5. Deploy
Next
Best
Action
4. Visualize
Dashboard

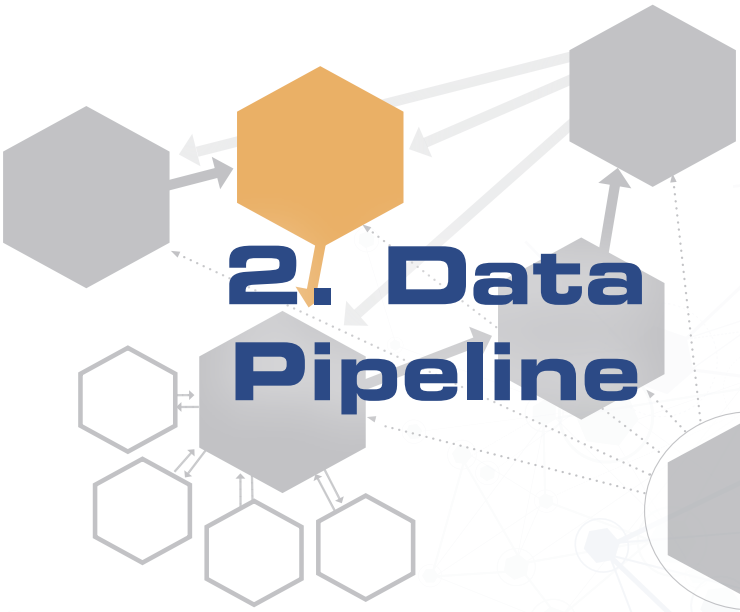


1. Data Source Integration

Data can be collected from various sources, both internal and external. Structured CRM data can be joined with free text sentiment data from social media. Traditional enterprise data warehouses can be integrated with unformatted new data sources using an enterprise data hub strategy. All data is then centralized in a data lake which usually involves a hadoop infrastructure in a cloud environment.

Much of this step relies on Big Data technologies, which provides almost unlimited data capacity and ingestion speed, schema-less data stores, stream processing, built in redundancy, auto-indexing, and parallel computing functions.

The resulting infrastructure easily hosts petabytes to even exabytes of data and readily serves them online to feed the requirements of the Machine Learning stage downstream.



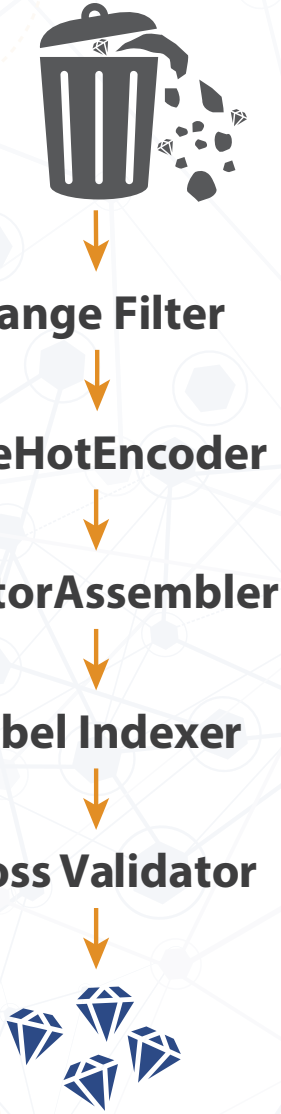
2. Data Pipeline

Disparate data in their raw form comes in with all kinds of gory inadequacies: missing data, typos, redundancies, erroneous schemas, incompatible formats, messy versioning, lost lineage, misinformation, sparsity, overwhelming volume, limited bandwidth, etc. Extracting the few gems from a large pile of useless garbage data in a repeatable operational setting necessitates the creation of a data pipeline. The system then cleanses out the noise and then package the result into a suitable form for applying mathematical algorithms.

Typically, feature reduction by means of Correlation Feature Selection (CFS) is done to cut down the raw data set to a manageable size. Other times, it may be desirable to extract new feature sets (indexing, aggregation, concatenation, etc.) that are not already present in the raw data. Sound feature engineering is frequently credited for winning machine learning contests where the algo-

rithms used by contestants are more or less indistinguishable.

While Big Data is credited for triggering the revival of Machine Learning, there are use cases where insufficient data is the main challenge. In those cases, the data pipeline performs the opposite function of supplying additional training data rather than filtering the initial set. There are well known statistical techniques for dealing with limited data such as bootstrapping, k-fold, and, where applicable, simulation from well known distributions. In cases of missing data, imputation techniques such as predictive mean matching, generalized low ranking model, or simple interpolation fills the gap. An effective data pipeline is therefore the combination of solid engineering and rigorous statistics.



In cases of missing data, imputation techniques such as predictive mean matching, generalized low ranking model, or simple interpolation fill the gap. An effective data pipeline is therefore the combination of solid engineering and rigorous statistics.

3. Run Hypothesis

Selection of the most appropriate predictive model is based on the input data volume, feature set, desired outcome, inclusion of time series, dimensionality and interpretability of the problem, and complexity of the mathematical formulation. Commonplace techniques such as linear regression and decision tree are well known and easily accessible from open source libraries. By and large, these techniques will solve many straightforward problems if applied correctly.

Real-world problems, however, have a way of quickly becoming chaotic. Multivariate Time series problems, for example, present sequential, vectorized datasets which

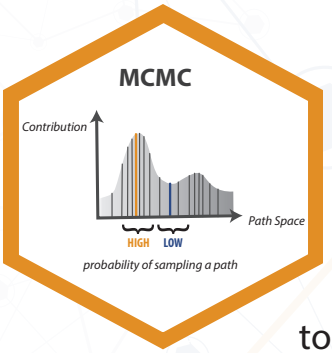
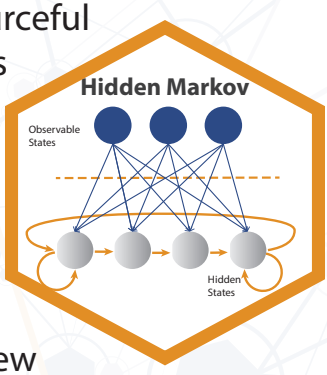


exhibit time behaviors not representable by static feature sets. Much of human decision making involves hidden states that are only weakly manifested by observable events. Complex problems such as symptom to disease mapping, or shopper funnel states require advanced machine learning techniques like Latent Classification Analysis and Hidden Markov Model, which deal specifically with these invisible states. On the extreme end of complexity, the phenomenon being studied by could involve dynamic, trend altering stimuli, target heterogeneous consumer groups, and rely on unnormalized survey responses and multiple layers of mutually dependent factors, leading to an absurdly unwieldy likelihood function, which can only be solved using the magic of Markov Chain Monte Carlo (MCMC). Fortunately, years of devoted research in Bayesian Inference provides a resourceful Data Scientist plenty of supporting tools (think in terms of inverse Wishart distribution, Gibbs Sampler library, etc.) to tackle these profoundly perplexing problems with impressive success.



The miraculous progress in Deep Learning technology, in the recent few years, opens up a brand new world of possibilities in applying machines to mimic human cognitive tasks. Everything that a human can process in under 1 second can now be done by Neural Networks with higher accuracy, in many cases, than the average (non-special-

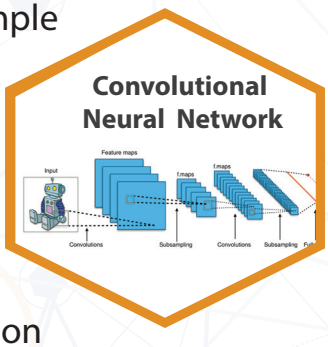
ist) person. As of 2015, machine vision recognizes images more accurately than humans. Similar benchmarks are achieved in speech and handwriting recognition. The two dominant Neural Network architectures, Convolution Neural Network (CNN) and Recurrent Neural Network (RNN) are being combined to perform the next level of AI feat, such as image captioning (describing a test image in complete sentences) and video recognition (identifying not only the objects in one image, but the motion depicted over consecutive frames of video, such as a baseball star hitting a home run). Although an overkill in some use cases, Deep Learning is generally the superior approach for problems with either a large sample size (> 1 million), or where complex relationships exist between the predictor variables.

Today, the holy grail of Deep Learning is called end-to-end learning, where the laborious and often trial-and -error step of feature engineering is summarily eliminated. This significant advancement both expands Neural Networks' range of applicability, and brings the their capabilities a huge step closer to the human level.

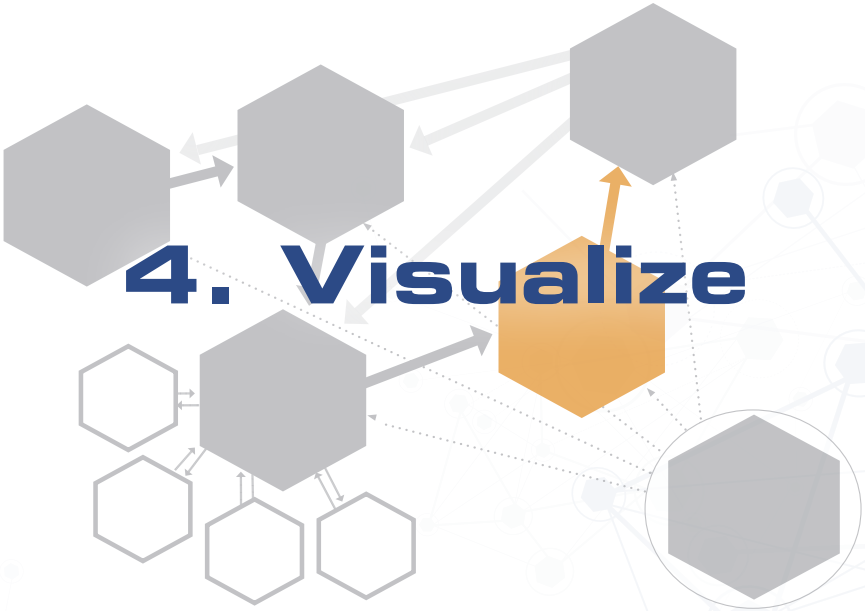
The process of selecting the technique to use for a particular use case is one which requires experience and deliberation. However, the real sweat is still to come after making the selection.

Needle in a Haystack

Data Science is often exalted as an exercise of the intellect. However, in many ways it is sheer test tube incubation. Overlooking one seemingly innocuous step can mean the difference between striking gold and collecting rubbish. Even the simple linear regression technique can fail from oscillating parameters due to spurious regression. Hidden Markov Model is susceptible to switching labels. All iterative convergence processes risk being trapped in a local minimum. Vector Autoregression requires a series of testings from Granger causality, to unit root, to co-integration tests. Overfitting is always a concern, whether optimizing a classic algorithm or training a Neural Network.



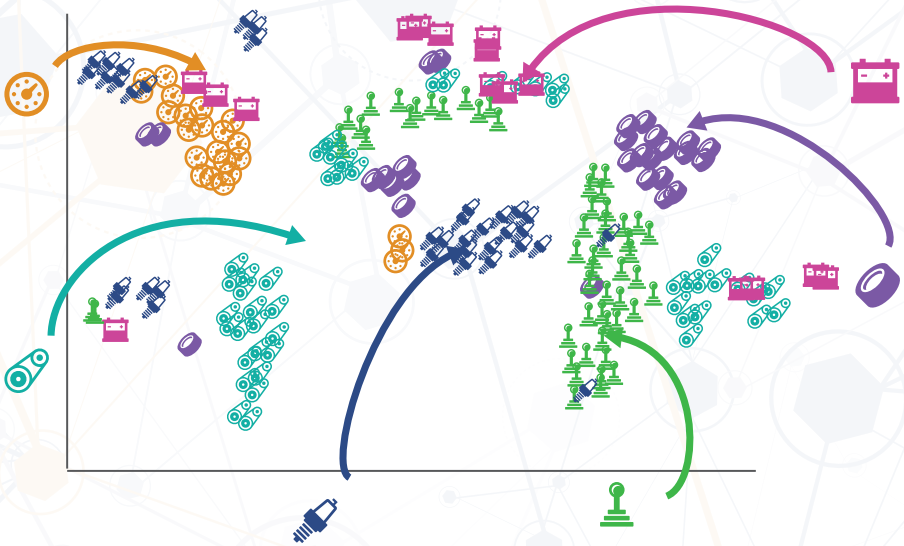
The long grind of Neural Network training does not mean leisurely workdays for the Data Scientists. It is a process of continuous tuning and midcourse corrections. Belated discovery of non-convergence translates to loss of precious time on expensive hardware. Slow convergence, high error rate, and systemic bugs are all potential signs of trouble that require corrective actions involving a wide array of parameters, including weight initialization, batch size, number of iterations, learning rate, activation function, loss function, optimization algorithm,s regularization, drop rate, etc. In short, well-tuned Neural Network training is as much an art as a science.

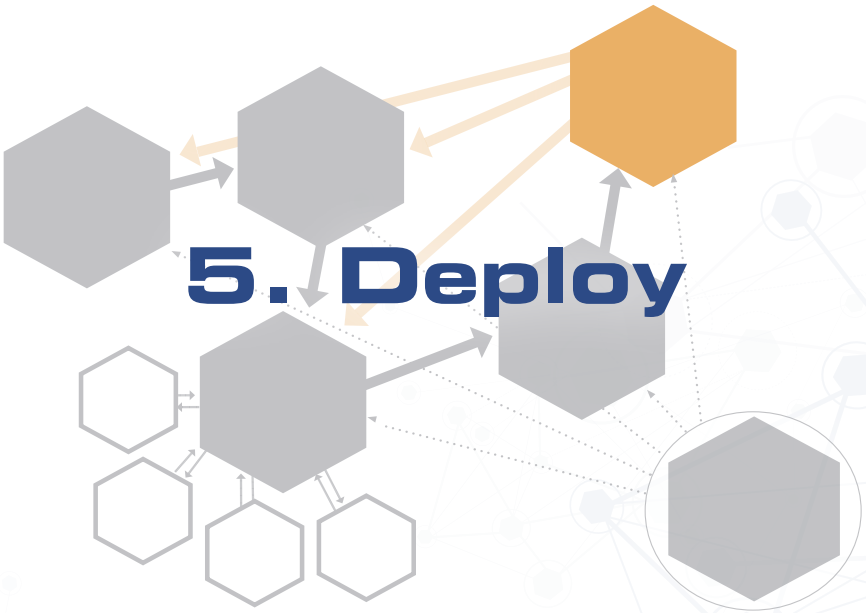


4. Visualize

Because of the wide ranging nature of business problems in the real world, and the chaotic state of available data, it is often necessary or advisable to first attain a level of understanding of the data itself before a committed investigation into making specific predictions. Michelangelo famously said every block of stone has a statue inside it and it is the task of the sculptor to discover it. In the same vein, there are hidden patterns in every collection of seemingly unwieldy data. The artisans who uncover the pattern reaps the information gold. The quickest way for making such a discovery is by means of visualization.

Recent advancements in data visualization takes the form of T-SNE with far reaching implications. Nearly boundless data, textual, numeric, or otherwise, can be represented as high-dimension vectors through a process called embedding. These high-dimensional vectors can then be automatically reduced to 2 or 3 dimensions for the purpose of visualization by a subject matter expert. With T-SNE, animal images can be grouped by species, words by semantic context, machine data by their deterioration behavior, and consumers by their purchase preferences. Analysts can now discover cause-and-effect relationships which directly result in improved marketing campaigns, design choices, investment strategies, etc.





5. Deploy

The training step results in a predictive model that can be used for making business decisions in production. A trained model is captured in a condensed bag of parameters along with a small footprint codebase reflecting the underlying Neural Network architecture. They can be deployed into a company's operational environment, generally after some level of system integration efforts. The predictive model feeds on real time market data and generates specific actionable advisories, with which front line managers can carry out operational tasks, be it bidding on real-time ads, targeting customers with personalized messages, taking preventative actions to discourage churns, or triggering supply chain adjustments.

The deployment stage must be concerned with performance requirements, due to the real-time nature of many use cases. However, since inferencing from a trained model is always highly efficient, even stringent performance requirements can usually be met fairly easily. Additional effort is needed to accelerate the model refresh process for making periodic updates to make sure the model remains accurate, as market conditions evolve. An automated back-testing infrastructure is needed in a similar vein to the continuous integration concept in software engineering.

Dashboard

The integrated system runs off a comprehensive dashboard, where operators monitor the predictive model traffic and throughput, performance metrics indicative of accuracy, and actual vs. expected impact, etc. Of particular importance is the monitoring of model accuracy, in a historical context. Although, every model has a refresh schedule, whereby retraining takes place using the latest data, sudden market changes may dictate an accelerated schedule, or even revamping of the model architecture. As much as business executives constantly react to shifting business climates, a predictive model is valid within a finite horizon. Frequent comparisons of current and past performances is a key to determining the right moment for making adjustments.

A Dashboard is the command center for predictive analytics monitoring. Non-data science staff conduct daily business and marketing operations, with the support of dashboard displays. The dashboard function must include the ability to run A/B tests to verify the validity of newly trained models. Additionally, data analysts responsible for the continuous workflow of the data pipeline should be able to manage ingestion, quality check, filtering, and finally the scheduled retraining of new models through this unified portal.


LOREM IPSUM
1,955
458
79
28
%25
3,155
1,283
JAN
FEB
MAR
APRIL
MAY
JUN
JULY
NAM RUTRUM
AENEAN

About BigR.io

BigR.io is a technology consulting firm empowering data to drive innovation and advanced analytics. We specialize in cutting-edge Big Data, Machine Learning and custom software strategy, analysis, architecture, and implementation solutions.

With extensive domain knowledge, BigR.io's architects and engineers design/build best-in-class solutions across a variety of verticals. This diverse industry exposure and our constant run-in with the cutting edge, arms us with invaluable tools, tricks, and techniques. Our knowledge and horsepower bring innovative, cost-conscious, and extensible results to complex software and data challenges.

BigR.io
EMPOWERING DATA

About Bruce Ho, Ph.D.

Bruce is a top-rated problem solver, communicator, and data scientist. Bruce has a unique strength in bridging software engineering with mathematical algorithms. He has over 15 years of IT experience, engaging in high-visibility, large-scale projects in marketing automation, cloud solutions, Big Data engineering and predictive analytics for top technology companies like Amazon, TeraData, and Life Technologies. He is a certified AWS solutions Architect, with a specialty in Big Data practices. Bruce published over 100 scientific papers during his academic career, and continues to draw the best ideas from state-of-the-art research, to fuel his Data Science practice.

Bruce has a deep-rooted passion in using statistics modeling to improve the way business is done in the fashion of "Moneyball". His fascination with the use of mathematics, to solve real world problems, dates back to the early days of Vector Calculus and Newtonian physics. After witnessing the growth of computing power in the mode of Moore's law, Bruce fully embraced the vision that data-driven decision making will soon become a basic tenet of business practice. He loves playing the advocate-in-chief for predictive analysis, when resolving client companies' challenges.

Bruce holds a BS in Physics from MIT, and an MS in Electrical Engineering & PhD in Applied Physics from Caltech.

One Boston Place, Suite #2600
Boston, MA 02108

617-500-5093 **info@bigr.io**